A Training-Based Mutual Information Lower Bound for Large-Scale Systems

Kang Gao, Xiangbo Meng, *Graduate Student Member, IEEE*, J. Nicholas Laneman, *Fellow, IEEE*, Jonathan D. Chisum, *Senior Member, IEEE*, Ralf Bendlin, *Senior Member, IEEE*, Aditya Chopra, *Senior Member, IEEE*, and Bertrand M. Hochwald, *Fellow, IEEE*

***Abstract*— We provide a mutual information lower bound that can be used to analyze the effect of training in models with unknown parameters. For large-scale systems, we show that this bound can be calculated using the difference between two derivatives of a conditional entropy function. We provide a step-by-step process for computing the bound, and apply the steps to a quantized large-scale multiple-antenna wireless communication system with an unknown channel. Numerical results demonstrate the interplay between quantization and training.**

***Index Terms*— Information rates, training, entropy, large-scale systems.**

I. Introduction

MANY systems have unknown parameters that are estimated during a training-phase with the help of known prescribed training signals. This phase is followed by a data phase, where knowledge of the estimated parameters is used to process the data. It is generally assumed that the parameters are constant during these two phases, the total duration of which is called the coherence time. It is often of great interest to optimize the training time for a given coherence time, since time in the training phase, while useful for parameter estimation, generally takes away from time in the data phase.

In a communication system, the parameters of interest often include the channel, which is typically unknown and learned at the receiver with the help of pilot signals sent by the transmitter. For example, [1] analyzes a multi-antenna model where a capacity lower-bound is obtained by using the minimum mean-square error (MMSE) estimate of the channel, and the residual channel error is treated as Gaussian noise. This lower bound is maximized over various parameters, including the fraction of the coherence time that should be dedicated to training. A similar optimization is considered in [2], where the power allocation and training duration are chosen to achieve the maximum sum-rate in a multiuser system. Such "one-shot learning," where the parameters are estimated only during the training phase, can be augmented by further refinement during the data phase [3], [4]. However, this refinement can suffer from error propagation [5], and we do not consider this herein.

We develop a framework to analyze one-shot training that does not require the parameters to appear linearly in the model, nor does it require additive Gaussian noise; rather, it requires the system to be time-invariant and memoryless, and a certain entropy to be computed in the large-scale system limit. This differentiates us from the previous efforts to analyze training, which assume that the unknown parameters appear linearly in the system model [1]–[4], [6], or appear in a linearized version of the model [7], [8], often by employing the Bussgang decomposition [9]. Herein, large-scale refers to infinite block lengths (time duration) or infinite-dimensional inputs and outputs, or both. The fact that the large-scale system entropy can sometimes be computed even when the small-scale system entropy cannot is exploited for our training analysis.

To demonstrate how to apply the developed framework, we consider a quantized large-scale multiple-input-multiple-output (MIMO) communication system, where the optimal training time and the corresponding data rate are investigated. By leveraging the large-scale results on certain entropies in [10]–[12], we are able to derive expressions for the data rate with training symbols and study the relationship between the optimal training time and various design parameters, including the ratio between the numbers of transmit and receive elements, the ratio between the blocklength and the number of transmit elements, and the number of resolution bits at both transmitters and receivers. Various conclusions on the interplay between training and the number of bits of resolution at the transmitter and receiver are provided.

This work is organized as follows. We first summarize some traditional methods to derive optimal amounts of training; then

Manuscript received 27 July 2021; revised 25 January 2022 and 18 March 2022; accepted 1 June 2022. Date of publication 13 June 2022; date of current version 16 August 2022. This work was generously supported by NSF Grant #1731056, and AT&T Labs. An earlier version of this paper was presented in part at ITA 2019, San Diego; and in part at Globecom Workshop 2019, HI, USA [DOI: 10.1109/GCWkshps45667.2019.9024647]. The associate editor coordinating the review of this article and approving it for publication was D. Tuninetti. *(Corresponding author: Xiangbo Meng.)*

Kang Gao was with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA. He is now with Qualcomm Technologies Inc., San Diego, CA 92121 USA (e-mail: kanggao@qti.qualcomm.com).

Xiangbo Meng, J. Nicholas Laneman, Jonathan D. Chisum, and Bertrand M. Hochwald are with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: xmeng@nd.edu; jnl@nd.edu; jchisum@nd.edu; bhochwald@nd.edu).

Ralf Bendlin is with AT&T Labs, Austin, TX 78712 USA (e-mail: rb691m@att.com).

Aditya Chopra was with AT&T Labs, Austin, TX 78712 USA. He is now with Amazon, Austin, TX 78759 USA (e-mail: adchopra@amazon.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2022.3182747.

Digital Object Identifier 10.1109/TCOMM.2022.3182747

in Section II, the problem for large-scale systems is formulated and the main results for computing the mutual information lower bound are stated; in Section III, the results developed in Section II are applied to a large-scale quantized MIMO communication system; Section IV concludes.

A. Brief Background and Prior Work

Consider a system model that has input and output processes $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots)$ and $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots)$, which comprise vectors $\mathbf{x}_t$ and $\mathbf{y}_t$ whose dimensions are M and N respectively. The input and output are connected through a conditional distribution parameterized by a vector or matrix G, whose value is unknown and whose size can be a function of T. We assume that G is constant during a coherence time block with length T, and then changes independently in the next block (same length T), and so on. The system is supplied with known inputs during a "training phase" to learn the parameters, after which the system is used during its "data phase." The unknown parameters are assumed to have a known distribution, and the number of unknown parameters is allowed to be a function of T.

A classical way to analyze the effects of training computes a lower bound on the mutual information between the input and output, as in [1], [2], [7], [8],

$$\frac{1}{T} I(\mathbf{x}^T; \mathbf{y}^T) \geq \frac{T - \tau T}{T} I(\mathbf{x}_{\tau T+1}; \mathbf{y}_{\tau T+1} | \mathbf{x}^{\tau T}, \mathbf{y}^{\tau T}), \quad (1)$$

where $\mathbf{x}_t$ and $\mathbf{y}_t$ are the tth vector input and output of the system, $\mathbf{x}^t = [\mathbf{x}_1^\intercal, \cdots, \mathbf{x}_t^\intercal]^\intercal$ collects all of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ into one long vector, $(\cdot)^\intercal$ denotes "transpose," $\mathbf{y}^t = [\mathbf{y}_1^\intercal, \cdots, \mathbf{y}_t^\intercal]^\intercal$, and τT is the number of training symbols in one coherence block. We assume $0 \leq \tau < 1$ is the fraction of the blocklength devoted to training, and τT is integer for convenience. (We choose this in favor of using $\lceil \tau T \rceil$ throughout.) The case $\tau = 0$ is interpreted as having no training; in other words, $\mathbf{x}^{\tau T}$ and $\mathbf{y}^{\tau T}$ are empty. The optimal training fraction

$$\tau_{\text{opt}} = \underset{\tau}{\text{argmax}}\ (1 - \tau) I(\mathbf{x}_{\tau T+1}; \mathbf{y}_{\tau T+1} | \mathbf{x}^{\tau T}, \mathbf{y}^{\tau T}) \quad (2)$$

then maximizes the lower bound (1). Although such analysis is frequently used, the right-hand side of (1) can be difficult to compute and is itself often approximated or lower bounded. For example, in [1], a wireless communication system with Rayleigh block-fading channel and additive Gaussian noise is considered, and the mutual information in (1) is lower bounded by treating the estimation error of the MMSE estimate of the channel as independent additive Gaussian noise. However, this form of analysis is often intractable when the parameters appear nonlinearly, or the additive noise is non-Gaussian, and explicit estimates of the unknown parameters are unavailable.

We leverage the fact that, sometimes, even with nonlinear models, the *large-scale* limits of the conditional entropies used to compute (1) are tractable. This provides us with a method to solve (2) for a large-scale system. In the next section, we provide the problem statement and main results.

II. Problem Statement and Main Results

A. Problem Statement

We reframe (2) as a large-scale limit problem. Let $T \to \infty$, and define the ratios

$$\alpha = \frac{N}{M}, \quad \beta = \frac{T}{M}. \quad (3)$$

It is possible, although not required, that M and N (the dimensions of the input and output) also grow to infinity with T, so that β is finite. The large-scale limit of the conditional mutual information $I(\mathbf{x}_{\tau T+1}; \mathbf{y}_{\tau T+1} | \mathbf{x}^{\tau T}, \mathbf{y}^{\tau T})$ in (2) is

$$\mathcal{I}'(\mathcal{X}; \mathcal{Y}) = \lim_{T \to \infty} \frac{1}{N} I(\mathbf{x}_{\tau T+1}; \mathbf{y}_{\tau T+1} | \mathbf{x}^{\tau T}, \mathbf{y}^{\tau T}). \quad (4)$$

The normalization by $1/N$ is needed to keep this quantity finite if $N \to \infty$, and this limit (assuming that it exists) typically depends on α, β, and τ. The optimal training time in (2) then becomes

$$\tau_{\text{opt}} = \underset{\tau}{\text{argmax}}\ (1 - \tau) \mathcal{I}'(\mathcal{X}; \mathcal{Y}). \quad (5)$$

We wish to solve (5). The value of this analysis depends on our ability to compute $\mathcal{I}'(\mathcal{X}; \mathcal{Y})$, and we show that this quantity can be computed as the derivative of a certain entropy.

B. Assumptions and Definitions of Useful Quantities

We first make some assumptions and definitions that are used by the main results. The bound in (1) is determined by the distribution of the triple $(\mathbf{x}^T, \mathbf{y}^T, G)$, and we make the following assumption:

$$\text{A1: } p(\mathbf{y}^T | \mathbf{x}^T, G) = \prod_{t=1}^{T} p(\mathbf{y}_t | \mathbf{x}_t, G), \quad (6)$$

$$p(\mathbf{x}^T) = p(\mathbf{x}^{\tau T}) \prod_{t=\tau T+1}^{T} p(\mathbf{x}_t), \quad (7)$$

where $p(\mathbf{y}_t | \mathbf{x}_t, G)$ is a time-invariant conditional distribution for all $t = 1, 2, \ldots, T$ and $p(\mathbf{x}_t)$ is a time-invariant distribution for all $t = \tau T + 1, \tau T + 2, \ldots, T$.

Equation (6) says that the system is memoryless and time-invariant (given the input and parameters) and (7) says that the input $\mathbf{x}_t$'s are independent and identically distributed (*iid*), and independent of $\mathbf{x}^{\tau T}$ for all $t > \tau T$. We use $p(\mathbf{x}^{\tau T})$ and $p(\mathbf{x}_t)$ to denote the distributions of $\mathbf{x}^{\tau T}$ and $\mathbf{x}_t$, respectively. Under A1, the distributions of $(\mathbf{x}^T, \mathbf{y}^T, G)$ are described by the set of distributions

$$\mathcal{P}(T, \tau) = \{p(\mathbf{y}|\mathbf{x}, G), p(G), p(\mathbf{x}^{\tau T}), p(\mathbf{x}_{\tau T+1})\}. \quad (8)$$

These distributions are used to calculate all of the entropies and mutual informations throughout. The entropies and mutual informations are "ergodic" in the sense that they are averaged over independent realizations of G.

Define:

$$\mathcal{H}'(\mathcal{Y}_\varepsilon | \mathcal{X}_\delta) = \lim_{T \to \infty} \frac{1}{N} H(\mathbf{y}_{\varepsilon \tau T+1} | \mathbf{x}^{\delta \tau T}, \mathbf{y}^{\varepsilon \tau T}), \quad (9)$$

$$\mathcal{H}'(\mathcal{Y}_\varepsilon | \mathcal{X}_{\delta^+}) = \lim_{T \to \infty} \frac{1}{N} H(\mathbf{y}_{\varepsilon \tau T+1} | \mathbf{x}^{\delta \tau T+1}, \mathbf{y}^{\varepsilon \tau T}), \quad (10)$$

with $\delta, \varepsilon \in [0, \frac{1}{\tau})$, again assuming these limits exist. Here, H can refer to either discrete or continuous entropy. We treat $\delta\tau T, \varepsilon\tau T$ again as integers to avoid excessive use of the ceiling or floor notation. We drop the subscripts ε and δ in $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ and $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\delta^+})$ when $\varepsilon = 1$ or $\delta = 1$. For example, $\mathcal{H}'(\mathcal{Y}|\mathcal{X}) = \mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)|_{\delta=1,\varepsilon=1}$ and $\mathcal{H}'(\mathcal{Y}|\mathcal{X}_+) = \mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\delta^+})|_{\delta=1,\varepsilon=1}$. With (9) and (10), we further make the following assumptions:

$$\text{A2: } \mathcal{H}'(\mathcal{Y}|\mathcal{X}_+) = \lim_{\varepsilon \searrow 1} \mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+}), \tag{11}$$

$$\mathcal{H}'(\mathcal{Y}|\mathcal{X}) = \lim_{\varepsilon \searrow 1} \mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}). \tag{12}$$

Notice that $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X})$ in (12) is $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)|_{\delta=1}$, where δ is dropped since $\delta = 1$, which is effectively taking $\lim_{\delta \searrow 1}$ before $\lim_{\varepsilon \searrow 1}$.

Assumptions A1–A2 in (6), (7), (11), and (12), are important for the main result (Theorem 1). A1 is often met in practice for a memoryless and time-invariant system with *iid* input in the data phase that is independent of the input and output during training. However, we do not have a complete characterization of the processes $\mathcal{X}$ and $\mathcal{Y}$ that meet Assumption A2. Nevertheless, A2 may be verified on a case-by-case basis by examining $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+})$ and $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X})$ with $\varepsilon \geq 1$ using Corollary 1(b) in Appendix A.

Define

$$\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \lim_{T\to\infty} \frac{1}{N\tau T} H(\mathbf{y}^{\varepsilon\tau T}|\mathbf{x}^{\delta\tau T}), \tag{13}$$

$$\mathcal{I}'(\mathcal{X}_\varepsilon; \mathcal{Y}_\varepsilon) = \lim_{T\to\infty} \frac{1}{N} I(\mathbf{x}_{\varepsilon\tau T+1}; \mathbf{y}_{\varepsilon\tau T+1}|\mathbf{x}^{\varepsilon\tau T}, \mathbf{y}^{\varepsilon\tau T}), \tag{14}$$

$$\mathcal{I}(\mathcal{X}_\varepsilon; \mathcal{Y}_\varepsilon) = \lim_{T\to\infty} \frac{1}{NT} I(\mathbf{x}^T_{\varepsilon\tau T+1}; \mathbf{y}^T_{\varepsilon\tau T+1}|\mathbf{x}^{\varepsilon\tau T}, \mathbf{y}^{\varepsilon\tau T}), \tag{15}$$

where $\mathbf{x}_t^T = [\mathbf{x}_t^\intercal, \mathbf{x}_{t+1}^\intercal, \cdots, \mathbf{x}_T^\intercal]^\intercal$, and $\mathbf{y}_t^T = [\mathbf{y}_t^\intercal, \mathbf{y}_{t+1}^\intercal, \cdots, \mathbf{y}_T^\intercal]^\intercal$. Similarly, we drop the subscripts ε and δ in $\mathcal{I}'(\mathcal{X}_\varepsilon; \mathcal{Y}_\varepsilon)$, $\mathcal{I}(\mathcal{X}_\varepsilon; \mathcal{Y}_\varepsilon)$, and $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ when $\varepsilon = 1$ or $\delta = 1$. Observe that, in particular, $\mathcal{I}(\mathcal{X}_0; \mathcal{Y}_0)$ corresponds to the large-scale limit of the mutual information between the input and output as shown in the left-hand side of (1), the only difference being the extra normalization factor $1/N$ needed to keep the limit finite. Taking the limit of (1) yields

$$\mathcal{I}(\mathcal{X}_0; \mathcal{Y}_0) \geq (1-\tau)\mathcal{I}'(\mathcal{X}; \mathcal{Y}). \tag{16}$$

For this inequality to be useful, we need to be able to compute the right-hand side of (16) as a function of τ. A method to do this is one of our main results.

C. Main Results

Theorem 1: Under Assumptions A1 and A2,

$$\mathcal{I}'(\mathcal{X}; \mathcal{Y}) = \lim_{\varepsilon \searrow 1} \frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})}{\partial \varepsilon} - \lim_{\varepsilon \searrow 1} \frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\varepsilon)}{\partial \varepsilon}. \tag{17}$$

Proof: Please see Appendix A. □

Theorem 1 shows that the right-hand side of (16) can be calculated as a derivative using (17) as long as $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ is available. We describe the process of computing this.

Computation of $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$: An expression for $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ may be derived from $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ (where $\delta = 1$) when this latter quantity is available. In some cases $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ can be obtained through methods employed in statistical mechanics by treating the conditional entropy as free energy in a large-scale system. Free energy is a fundamental quantity [13], [14] that has been analyzed through the powerful "replica method," and this, in turn, has been applied to entropy calculations in machine learning [15]–[18] and wireless communications [10]–[12], in both linear and nonlinear systems.

The entropy $\mathcal{H}(\mathcal{Y}|\mathcal{X})$ (where $\delta = \varepsilon = 1$) is considered in [15]–[18], where the input is multiplied by an unknown vector as an inner product and then passes through a nonlinearity to generate a scalar output. In [15], [16], [18], the inputs are *iid*, while orthogonal inputs are considered in [17]. The entropy $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ for MIMO systems is considered in [10]–[12], where the inputs are *iid* in the training phase and are *iid* in the data phase, but the distributions in the two phases can differ. In [10], a linear system is considered where the output is the result of the input multiplied by an unknown matrix, plus additive noise, while in [11], [12] uniform quantization is added at the output.

As we now show, the expression for $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ for $\varepsilon \geq 1$ can be leveraged to compute $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ for all $\varepsilon, \delta > 0$. We consider the case when the input $\mathbf{x}_t$ is *iid* for all t, and the distribution set $\mathcal{P}(T, \tau)$ defined in (8) can therefore be simplified as

$$\mathcal{P}(T, \tau) = \{p(\mathbf{y}|\mathbf{x}, G), p(G), p(\mathbf{x})\}. \tag{18}$$

The following theorem assumes that we have $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ available as a function of τ for all $\varepsilon \geq 1$.

Theorem 2: Assume that Assumption A1 is met, $\mathbf{x}_t$ are iid for all t, $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ defined in (13) *exists and is continuous in $\tau \in (0, 1)$ and $\varepsilon \geq 1$. Define*

$$F(\tau, \varepsilon) = \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}). \tag{19}$$

Then

$$\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = u \cdot F\left(u\tau, \frac{\varepsilon - u}{\delta} + 1\right), \tag{20}$$

for all $\varepsilon, \delta \in (0, \frac{1}{\tau}]$, where $u = \min(\varepsilon, \delta)$.

Proof: According to (13) and (19), we have

$$F(\tau, \varepsilon) = \lim_{T\to\infty} \frac{1}{N\tau T} H(\mathbf{y}^{\varepsilon\tau T}|\mathbf{x}^{\tau T}),$$

which is computed using $\mathcal{P}(T, \tau)$ defined in (18). When $\delta \geq \varepsilon > 0$, we have

$$\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \lim_{T\to\infty} \frac{1}{N\tau T} H(\mathbf{y}^{\varepsilon\tau T}|\mathbf{x}^{\delta\tau T}) \tag{21}$$

$$= \lim_{T\to\infty} \frac{1}{N\tau T} H(\mathbf{y}^{\varepsilon\tau T}|\mathbf{x}^{\varepsilon\tau T})$$

$$= \lim_{T\to\infty} \frac{\varepsilon}{N\tilde{\tau} T} H(\mathbf{y}^{\tilde{\tau} T}|\mathbf{x}^{\tilde{\tau} T}). \tag{22}$$

where $\tilde{\tau} = \varepsilon\tau$. Therefore, (19) and (22) yield

$$\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \varepsilon \cdot F(\tilde{\tau}, 1) = \varepsilon \cdot F(\varepsilon\tau, 1). \tag{23}$$

When $\varepsilon > \delta > 0$, let $\tilde{\tau} = \delta\tau$, and then (21) yields

$$\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \lim_{T\to\infty} \frac{\delta}{N\tilde{\tau}T} H(\mathbf{y}^{\varepsilon\tilde{\tau}T/\delta}|\mathbf{x}^{\tilde{\tau}T}) = \delta \cdot F\left(\tilde{\tau}, \frac{\varepsilon}{\delta}\right) = \delta \cdot F\left(\delta\tau, \frac{\varepsilon}{\delta}\right). \quad (24)$$

By combining (23) and (24), we obtain (20). □

With Theorems 1 and 2, we may summarize the process of obtaining solving (5).

D. Steps for Computing Optimal Training Fraction

Summary: We assume that the input dimension M, the output dimension N, and the coherence time (block of symbols) T have the ratios defined in (3). The unknown parameters of the system are constant within the block, and change independently in the next block. The first τT symbols of each block are used for training and the remaining $T - \tau T$ are for data. The input $\mathbf{x}_t$ are *iid* for all $t = 1, \ldots, T$. Using results from Appendix A, the computation then follows these seven steps:

1) Verify Assumption A1 (6)–(7) based on the set of distributions in (8).
2) Compute $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ defined in (13) for $\varepsilon \geq 1$ and express it as a function of τ and ε, as in $F(\tau, \varepsilon)$ (19).
3) Compute $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ defined in (13), for all $\varepsilon, \delta \in (0, \frac{1}{\tau}]$ by using Theorem 2 and $F(\tau, \varepsilon)$.
4) Compute $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ and $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+})$ defined in (9)-(10) by taking the derivative of $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ and $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\varepsilon)$ (Theorem 4 and Corollary 1(a) in Appendix A).
5) Verify Assumption A2 (11) by examining the expressions of $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+})$ and verify (12) with Corollary 1(b) in Appendix A.
6) Compute $\mathcal{I}'(\mathcal{X}; \mathcal{Y})$ by using (17).
7) Solve τ_{opt} using (5).

The solution to (5) is then an approximation of (2).

The results of the above process are most useful when the number of unknowns scales with T. Otherwise, the results may be uninteresting. For example, consider a system modeled as

$$y_t = g x_t + v_t, \quad t = 1, 2, \ldots$$

where g is the unknown gain of the system, x_t, y_t are the input and corresponding output, v_t is the additive noise, τ is the fraction of time used for training. This system is bilinear in the gain and the input. We assume that v_t is modeled as *iid* Gaussian $\mathcal{N}(0, 1)$, independent of the input. The training signals are $x_t = 1$ for all $t = 1, 2, \ldots, \tau T$, and the data signals x_t are modeled as *iid* Gaussian $\mathcal{N}(0, 1)$ for all $t = \tau T + 1, \tau T + 2, \ldots$ An analysis that follows the steps above produces

$$\mathcal{I}(\mathcal{X}_0; \mathcal{Y}_0) \geq \frac{1-\tau}{2} \mathbb{E}_g \log(1 + g^2),$$

and therefore $\tau_{\text{opt}} = 0$ maximizes this bound. This result reflects the fact that g is learned perfectly for any $\tau > 0$ because there is only one unknown parameter for τT training symbols as $T \to \infty$. Hence, trivially, it is advantageous to make τ as small as possible.

The next section applies the results of this section to a quantized MIMO system where the number of unknown channel coefficients approaches infinity as $T \to \infty$.

III. Application to Large-Scale Quantized MIMO Systems

We consider a large-scale MIMO system [12], [19]–[22] modeled by

$$\mathbf{y}_t = \mathbf{f}\left(\sqrt{\frac{\rho}{M}} G\mathbf{x}_t + \mathbf{v}_t\right), \quad t = 1, 2, \ldots, \quad (25)$$

where $\mathbf{x}_t \in \mathcal{C}^{M\times 1}$ models the transmitted signals from M elements (transmitter antennas) at time t, $\mathbf{y}_t \in \mathcal{C}^{N\times 1}$ models the received signals with N elements (receiver antennas), $G \in \mathcal{C}^{N\times M}$ models the unknown baseband-equivalent wireless channel whose elements have *iid* real and imaginary components with zero-mean half-variance common distribution $p_{\tilde{g}}(\cdot)$, $\mathbf{v}_t \in \mathcal{C}^{N\times 1}$ models the additive white Gaussian noise at the receiver whose elements are *iid* circular-symmetric complex Gaussian $\mathcal{CN}(0, \sigma^2)$, and $\mathbf{f}(\cdot)$ is an element-wise function that applies b-bit uniform quantization $f(\cdot)$ to each element. The real and imaginary parts are quantized separately.

The system model includes a coherence blocklength T during which the channel is constant and after which it changes independently to a new value. We consider a Rayleigh environment, where the channel G has *iid* $\mathcal{CN}(0, 1)$ elements, and the corresponding $p_{\tilde{g}}(\cdot)$ is $\mathcal{N}(0, \frac{1}{2})$. A fraction τ of the total blocklength is used to learn the channel G by transmitting known training signals. We consider $N, M, T \to \infty$ with fixed ratios defined in (3) as an approximation for finite but large input/output dimensions or long blocklength.

We wish to determine the optimal training fraction τ_{opt} of such systems defined in (5), and the corresponding optimal achievable rate (in "bits/channel-use/transmitter")

$$\mathcal{R}_{\text{opt}} = \max_\tau (1 - \tau)\alpha \mathcal{I}'(\mathcal{X}; \mathcal{Y}), \quad (26)$$

by using the process developed in Section II-D. First, to compute $\mathcal{I}'(\mathcal{X}; \mathcal{Y})$, we make the following assumptions and define several useful functions.

A. Assumptions and Definitions of Useful Quantities

We assume that the real and imaginary elements of the transmitted vector $\mathbf{x}_t$ are *iid* with zero mean and half variance common distribution $p_{\tilde{x}}(\cdot)$ and are a-bit quantized signals in both the in-phase and the quadrature branches. This creates a 2^{2a}-QAM constellation, with all possible symbols generated with equal probability; the corresponding $p_{\tilde{x}}(\cdot)$ is uniform among the 2^a real and imaginary components. Throughout this section we assume $\sigma^2 = 1$, and thus the quantity ρ is nominally the signal-to-noise ratio (SNR) because it represents the ratio of the average signal energy $(\rho/M)\mathbb{E}\|G\mathbf{x}_t\|^2 = \rho N$ to noise variance $N\sigma^2 = N$, before quantization with the function $f(\cdot)$.

The b-bit quantizer $f(\cdot)$ has 2^b-1 real quantization thresholds defined as

$$r_k = (-2^{b-1}+k)\Delta, \quad \text{for } k = 1, 2, \cdots, 2^b-1, \tag{27}$$

where Δ is the quantization step size. We define $r_0 = -\infty$, and $r_{2^b} = +\infty$ for convenience. The output of the quantizer indicates the quantization level: $f(w) = k$ for $w \in (r_{k-1}, r_k]$ and $k = 1, \ldots, 2^b$. When the input to the quantizer is a complex number, its real and imaginary parts are quantized independently. It is assumed throughout our numerical results that Δ is chosen such that $f(w) = 1$ or $f(w) = 2^b$ with probability $1/2^b$ when the input distribution on w is real Gaussian with mean zero and variance $(1+\rho)/2$. This choice of Δ ensures that ADCs are operating in a reasonable range, but it only affects the numerical results in Section III-C.

In order to compute $\mathcal{I}'(\mathcal{X};\mathcal{Y})$, we need to define some useful quantities and functions: $I_{\text{AWGN}}(\lambda, p(\cdot))$, $\mathcal{E}(\lambda, p(\cdot))$, $\Omega(\gamma, s)$, and $\chi(\gamma, s)$. Let $I_{\text{AWGN}}(\lambda, p_{\tilde{x}}(\cdot))$ and $\mathcal{E}(\lambda, p_{\tilde{x}}(\cdot))$ be

$$\begin{aligned} I_{\text{AWGN}}(\lambda, p_{\tilde{x}}(\cdot)) &= -\mathbb{E}_y[\log_2 \mathbb{E}_x(e^{-|y-\sqrt{\lambda}x|^2})] \\ &\quad - \log_2 e, \end{aligned} \tag{28}$$

$$\mathcal{E}(\lambda, p_{\tilde{x}}(\cdot)) = \mathbb{E}_{x,y}(|x - \int x \cdot p(x|y)dx|^2), \tag{29}$$

where the expectation is with respect to the joint distribution of (x, y):

$$\begin{aligned} p(x,y) &= p(x) \cdot \frac{1}{\pi} e^{-|y-\sqrt{\lambda}x|^2} \\ &= p_{\tilde{x}}(x_{\text{R}}) \cdot p_{\tilde{x}}(x_{\text{I}}) \cdot \frac{1}{\pi} e^{-|y-\sqrt{\lambda}x|^2}, \end{aligned} \tag{30}$$

where x_{R} and x_{I} are the real and imaginary part of x and $p(x|y)$ is the distribution of x conditioned on y. Note that (x, y) can be modeled as a single-input single-output additive white Gaussian noise (AWGN) channel

$$y = \sqrt{\lambda}x + v, \tag{31}$$

where $v \sim \mathcal{CN}(0,1)$ is independent of x. Then $I_{\text{AWGN}}(\lambda, p_{\tilde{x}}(\cdot))$ is the mutual information between x and y, and $\mathcal{E}(\lambda, p_{\tilde{x}}(\cdot))$ is the mean-square error (MSE) of the MMSE estimate of x conditioned on y. We define $I_{\text{AWGN}}(\lambda, p_{\tilde{g}}(\cdot))$ and $\mathcal{E}(\lambda, p_{\tilde{g}}(\cdot))$ in a similar way by replacing $p_{\tilde{x}}(\cdot)$ in the above expressions with $p_{\tilde{g}}(\cdot)$.

We define $\Omega(\gamma, s)$ and $\chi(\gamma, s)$ as

$$\begin{aligned} \Omega(\gamma, s) &= -2\sum_{k=1}^{2^b} \int_{\mathcal{R}} \Psi_k(\sqrt{\gamma}z, s) \\ &\quad \times \log_2 \Psi_k(\sqrt{\gamma}z, s) \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz, \end{aligned} \tag{32}$$

$$\chi(\gamma, s) = \sum_{k=1}^{2^b} \int_{\mathcal{R}} dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \frac{(\Psi_k'(\sqrt{\gamma}z, s))^2}{\Psi_k(\sqrt{\gamma}z, s)}, \tag{33}$$

where

$$\begin{aligned} \Psi_k(w, s) &= \Phi\left(\frac{\sqrt{2}r_k - w}{\sqrt{s}}\right) \\ &\quad - \Phi\left(\frac{\sqrt{2}r_{k-1} - w}{\sqrt{s}}\right), \end{aligned} \tag{34}$$

$$\Psi_k'(w, s) = \frac{e^{-\frac{(\sqrt{2}r_k - w)^2}{2s}} - e^{-\frac{(\sqrt{2}r_{k-1} - w)^2}{2s}}}{\sqrt{2\pi s}}, \tag{35}$$

with r_k defined in (27), and $\Phi(\cdot)$ is the standard normal cumulative distribution function.

We use $a = \infty$ to denote an unquantized transmitter and assume the elements of $\mathbf{x}_t$ are *iid* $\mathcal{CN}(0,1)$; the corresponding $p_{\tilde{x}}(\cdot)$ is then $\mathcal{N}(0, \frac{1}{2})$. We use $b = \infty$ to denote $f(w) = w$ for $w \in \mathcal{R}$ (quantizer is removed). In this case, $I_{\text{AWGN}}(\lambda, p(\cdot))$, $\mathcal{E}(\lambda, p(\cdot))$ are as above, but $\Omega(\gamma, s)$ and $\chi(\gamma, s)$ are

$$\Omega(\gamma, s) = \log_2(\pi e s), \quad \chi(\gamma, s) = 1/s, \tag{36}$$

and the computed $\mathcal{H}'(\mathcal{Y}|\mathcal{X})$ and $\mathcal{H}'(\mathcal{Y}|\mathcal{X}_+)$ are then differential entropies.

We may now compute $\mathcal{I}'(\mathcal{X};\mathcal{Y})$ for the MIMO quantizer model.

B. $\mathcal{I}'(\mathcal{X};\mathcal{Y})$ for the System in Equation (25)

Theorem 3: For the system (25) *with τT input-output training pairs $(\mathbf{x}^{\tau T}, \mathbf{y}^{\tau T})$,*

$$\begin{aligned} \mathcal{I}'(\mathcal{X};\mathcal{Y}) &= \Omega(\rho q_g q_x, \sigma^2 + \rho - \rho q_g q_x) \\ &\quad - \Omega(\rho q_g, \sigma^2 + \rho - \rho q_g) \\ &\quad + \frac{1}{\alpha} I_{\text{AWGN}}(\lambda_x, p_{\tilde{x}}(\cdot)) + \frac{q_x \lambda_x - \lambda_x}{\alpha \ln 2}, \end{aligned} \tag{37}$$

where $\Omega(\cdot,\cdot)$ is defined in (32)*, $I_{\text{AWGN}}(\cdot,\cdot)$ is defined in* (28)*, $(q_g, \lambda_g, q_x, \lambda_x)$ are the solutions of*

$$\begin{aligned} \lambda_g &= \tau\beta\rho \cdot \chi(\rho q_g, \sigma^2 + \rho - \rho q_g), \\ q_g &= 1 - \mathcal{E}(\lambda_g, p_{\tilde{g}}(\cdot)), \end{aligned} \tag{38}$$

$$\begin{aligned} \lambda_x &= \alpha\rho q_g \cdot \chi(\rho q_g q_x, \sigma^2 + \rho - \rho q_g q_x), \\ q_x &= 1 - \mathcal{E}(\lambda_x, p_{\tilde{x}}(\cdot)), \end{aligned} \tag{39}$$

where $\mathcal{E}(\cdot,\cdot)$ and $\chi(\cdot,\cdot)$ are defined in (29) *and* (33)*. When $f(w) = w$, $\Omega(\cdot,\cdot)$ and $\chi(\cdot,\cdot)$ in* (32) *and* (33) *are replaced with* (36)*.*

Proof: The proof applies Theorems 1 and 2, and is detailed in Appendix B. □

Unlike the analysis in [1]–[4] or [6]–[8], the theorem: (i) provides a large-scale expression for $\mathcal{I}'(\mathcal{X};\mathcal{Y})$, not a lower bound; (ii) does not require worst-case noise analysis or linearization of the quantizer $f(\cdot)$.

Although we compute large-scale limits, it is anticipated that the results herein provide good approximations for systems with finite M, N, and T simply by substituting the α and β computed for the finite-dimensional system into the limiting formulas. For example, the parameter $\beta = T/M$ is the ratio of the coherence time of the channel (in symbols) to the number of transmitters and is therefore strongly dependent on the physical environment. We may choose a typical value

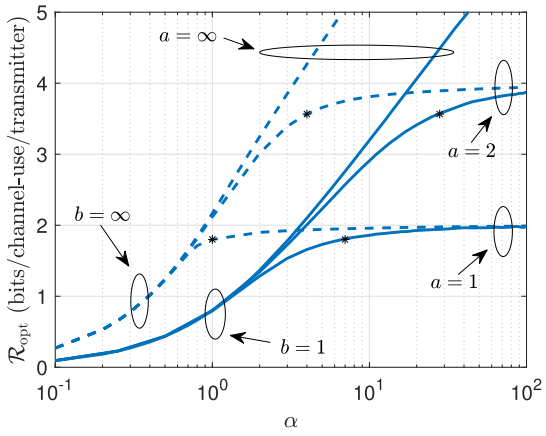


Fig. 1. Plots of $\mathcal{R}_{\rm opt}$ vs α for $a=1,2,\infty$, and $b=1$ (solid curves) and $b=\infty$ (dashed curves), and SNR $=10$ dB, for $\beta=40$ (see text for explanation of this choice). The curves saturate at 2 for $a=1$, and at 4 for $a=2$, while there is no saturation for $a=\infty$. Note for $\mathcal{R}_{\rm opt}$ to achieve the 90% level ($1.8a$ indicated by "*"), more receive elements are required when comparing linear ($b=\infty$) versus one-bit ($b=1$) receivers. For example, with $a=2$, observe that the "*" on the dashed curve is at $\alpha=4$, while the "*" on the solid curve is at $\alpha=28$, indicating that α has to increase to compensate for lack of resolution at the receiver.

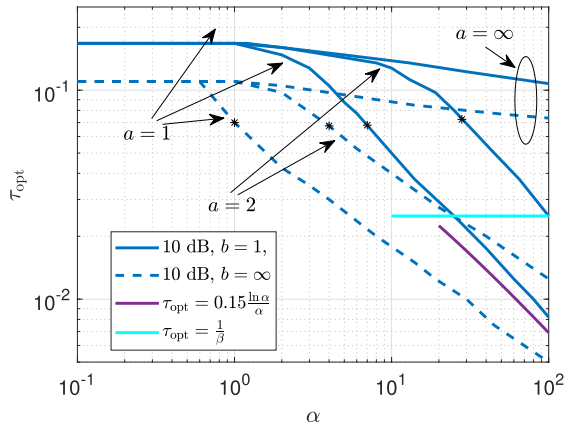


Fig. 2. Plots of $\tau_{\rm opt}$ vs α. Note that $\tau_{\rm opt}$ is generally insensitive to α when α is small, and decreases rapidly with α as $\mathcal{R}_{\rm opt}$ approaches the saturation rate $2a$ (after the markers, which indicate $\mathcal{R}_{\rm opt}=1.8a$, same as in Fig. 1). As α grows, eventually $\tau_{\rm opt}\cdot\beta<1$ (indicated by the solid cyan line), at which point the number of training symbols is smaller than the number of transmitter elements. Also shown in purple is the large-α result (42) for 10 dB SNR, $a=1$ and $b=1$.

as follows: Suppose we choose a 3.5 GHz carrier frequency with maximum mobility of 130 km/h; the maximum Doppler shift becomes $f_{\rm d}=\frac{130\text{ km/h}\times 3.5\text{ GHz}}{3\times 10^8\text{ m/s}}=421$ Hz, and the corresponding coherence time is $\frac{9}{16\pi f_{\rm d}}=0.4$ ms [23]. We consider 10 MHz bandwidth and assume that the system is operated at Nyquist sampling rate (10 complex Msamples/second), which produces $T=4000$ discrete samples during each 0.4 ms coherent block. In a system with $M=100$ elements at the transmitter, we obtain $\beta=40$.

With the expression (37), we are now able to study the effects of quantization on training.

C. Optimal Training Time and Achievable Rate for Various Scenarios

In this section, we consider the results of (5) and (26) for various scenarios, where we show how the optimal training time and the achievable rate is affected by quantization (here quantization means a, or b, or both are finite). Because of quantization, the maximum achievable rate is $2a$ bits/channel-use/transmitter, which we define to be the saturation rate. We note that the large-scale results derived below can be used for finite systems. For example, to determine the optimal number of training symbols, we first solve the large-scale problem (5) to get the optimal training fraction $\tau_{\rm opt}$, and then the number of training symbols is just $\tau_{\rm opt}T$.

1) Lower Resolution at Receiver Requires More Receiver Elements: In Fig. 1 we consider $\mathcal{R}_{\rm opt}$ versus α for $a=1,2$; the maximum rates per transmitter are then 2 bits and 4 bits respectively ($2a$ bits/channel-use/transmitter is considered as the saturation rate). These asymptotes are approached as α is increased, where larger $\alpha=N/M$ represents larger number of receivers per transmitter. The unquantized results are also shown for comparison. For example, by comparing the solid curves ($b=1$) versus the dashed curves ($b=\infty$), we can see that to achieve the same rate, larger α is needed for ($b=1$) which indicates that more receiver elements are required to compensate for the lack of resolution.

2) $\tau_{\rm opt}T<M$ Is Sufficient for Quantized Systems When α Is Large: In Fig. 2, we show that for $a=1,2$, when α is large, the optimum number of training signals may be smaller than the number of transmitters. Hence, unlike a linear system where an explicit estimate of H is learned only if $\tau_{\rm opt}T\geq M$, it is not necessary to learn all of the $M\cdot N$ channel coefficients in a quantized system when $N\gg M$. This is also shown to a limited extent for $a=1$ and $b=1$ in [24], [25]. We further show that, for all finite a, the optimum training time decreases to zeros as α increases to infinity; especially, $\tau_{\rm opt}$ decays as $(\ln\alpha)/\alpha$ for large α when $a=1$.

Specifically, for all $\tau>0$ and finite a, we have

$$\lim_{\alpha\to\infty}\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})=2a, \tag{40}$$

which yields $\lim_{\alpha\to\infty}\mathcal{R}_{\rm opt}=2a$, and $\lim_{\alpha\to\infty}\tau_{\rm opt}=0$. The proof is shown in Appendix C. In the special case with $a=1$, when α is large, we have

$$\text{for } b=\infty,\quad \tau_{\rm opt}\approx 2\left(\frac{\rho+1}{\rho}\right)^2\frac{\ln\alpha}{\beta\alpha}, \tag{41}$$

$$\text{for } b=1,\quad \tau_{\rm opt}\approx 2\left(\frac{\pi}{2}\frac{\rho+1}{\rho}\right)^2\frac{\ln\alpha}{\beta\alpha}. \tag{42}$$

3) Linearization Can Sometimes Work Well at High SNR: It is well-known that linearization at the receiver works well at low SNR's [7], [26]. We show that it sometimes also works well at high SNR with quantized transmitters (a is finite). By using the Bussgang decomposition [9], [26], we can reformulate the system in (25) as

$$\mathbf{y}_t=\sqrt{\frac{\eta}{(\rho+1)}}\left(\sqrt{\frac{\rho}{M}}G\mathbf{x}_t+\mathbf{v}_t\right)+\mathbf{v}_{\rm q}, \tag{43}$$

where $\mathbf{v}_{\rm q}$ is uncorrelated with $\sqrt{\frac{\rho}{M}}G\mathbf{x}_t + \mathbf{v}_t$, $\mathbf{v}_{\rm q}$ has zero mean with covariance matrix $(1-\eta)I$, and where $\eta = 2/\pi$ for $b=1$, and $\eta = \frac{2}{5\pi}(1+2e^{-\frac{\Delta^2}{\rho+1}})^2$ for $b=2$. For tractability, we assume that $\mathbf{v}_{\rm q} \sim \mathcal{CN}(0,(1-\eta)I)$ and is independent of $G, \mathbf{x}_t$, and $\mathbf{v}_t$.

The classical treatment of this model treats the estimated channel as the "true" channel as in [1], while the estimation error is treated as additive Gaussian noise, thereby furthering the approximation. We obtain

$$\bar{\mathbf{y}}_t = \sqrt{\rho_{\rm eff}/M}\bar{G}\mathbf{x}_t + \bar{\mathbf{v}}_t, \tag{44}$$

where $\rho_{\rm eff}$ is the effective SNR

$$\rho_{\rm eff} = \frac{\tau\beta\eta^2\rho^2}{[1+(1-\eta)\rho][1+\rho+\tau\beta\eta\rho]}, \tag{45}$$

$\bar{G}$ is the estimated channel whose elements are *iid* $\mathcal{CN}(0,1)$, and $\bar{\mathbf{v}}_t$ has *iid* $\mathcal{CN}(0,1)$ elements. This model has achievable rate

$$\lim_{M\to\infty}\frac{1}{M}I(\mathbf{x};\bar{\mathbf{y}}|\bar{G}) = \alpha\log_2(1+\rho_{\rm eff}-\rho_{\rm eff}\bar{q}_x) + I_{\rm AWGN}(\bar{\lambda}_x, p_{\tilde{x}}(\cdot)) + \frac{\bar{q}_x\bar{\lambda}_x - \bar{\lambda}_x}{\ln 2}, \tag{46}$$

where $(\bar{q}_x, \bar{\lambda}_x)$ are the solutions of

$$\begin{aligned}\bar{\lambda}_x &= \alpha\rho_{\rm eff}\cdot\chi(\rho_{\rm eff}\bar{q}_x, 1+\rho_{\rm eff}-\rho_{\rm eff}\bar{q}_x),\\ \bar{q}_x &= 1-\mathcal{E}(\bar{\lambda}_x, p_{\tilde{x}}(\cdot)).\end{aligned} \tag{47}$$

The details for computing (46) are shown in Appendix D. Note that $\lim_{M\to\infty}\frac{1}{M}I(\mathbf{x};\bar{\mathbf{y}}|\bar{G})$ is a function of τ through $\rho_{\rm eff}$ in (45). We then define

$$\mathcal{R}_{\rm L} = \max_{\tau}(1-\tau)\lim_{M\to\infty}\frac{1}{M}I(\mathbf{x};\bar{\mathbf{y}}|\bar{G}). \tag{48}$$

The path just described to obtain $\mathcal{R}_{\rm L}$ involves several approximations, and hence it is unclear how closely $\mathcal{R}_{\rm L}$ should follow $\mathcal{R}_{\rm opt}$. However, a comparison between $\mathcal{R}_{\rm opt}$ (26) and $\mathcal{R}_{\rm L}$ (48) with $a=1,2,\infty$ and $b=1,2$ for $\beta=40$ is shown in Fig. 3 with $\alpha=10$. We can see that $\mathcal{R}_{\rm L}$ is generally a good approximation of $\mathcal{R}_{\rm opt}$ when the SNR is below 6 dB, but is also accurate above 6 dB in cases where $\mathcal{R}_{\rm opt}\approx 2a$ (saturation rate) when SNR $\approx$ 6 dB; see especially the blue, black, and green curves in Fig. 3.

4) Minimum Number of Receivers per Transmitter Is Required for Quantized Systems: The values of α required to achieve $\mathcal{R}_{\rm opt} = 1.8$ (90% level) for various SNR ρ and b with $a=1$ are shown in Fig. 4. It is clear that α decreases as ρ increases, but there are asymptotes when $\rho=\infty$ for $b=1,2,3$ as shown in the cyan lines because of the quantization noise. This indicates that for quantized systems, if α is below some minimum (cyan lines), no matter how large the SNR is, certain data rates are not achievable. However, when $b=\infty$, there is no asymptote since the channel can be estimated perfectly, and the discrete transmitted signal can be detected perfectly as $\rho\to\infty$.

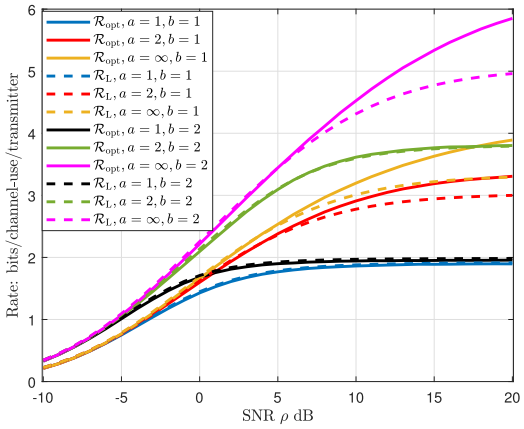


Fig. 3. Plots of $\mathcal{R}_{\rm opt}$ and $\mathcal{R}_{\rm L}$ vs SNR with $\beta=40$ and $\alpha=10$ for $a=1,2,\infty$ and $b=1,2$. Observe that $\mathcal{R}_{\rm L}$ is a good approximation of $\mathcal{R}_{\rm opt}$ below 6 dB SNR, and is sometimes also a good approximation for all SNR, depending on where saturation (rate $2a$) is reached.

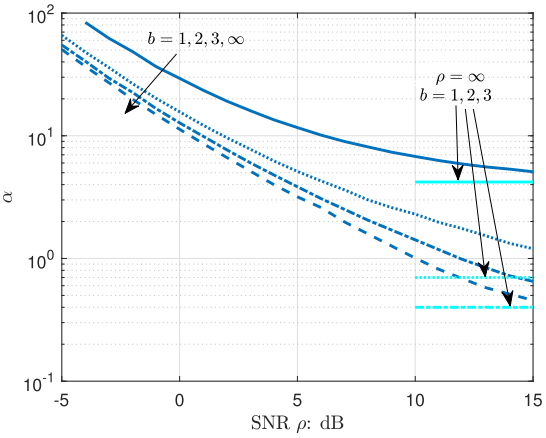


Fig. 4. Plots of α versus SNR when $\mathcal{R}_{\rm opt}=1.8$ (90% level), $\beta=40$, $a=1$. The curves decrease as SNR increases and reach asymptotes as $\rho=\infty$ that are shown in cyan with $b=1,2$, and 3, which indicates the minimum number of receivers per transmitter required to maintain $\mathcal{R}_{\rm opt}$. These asymptotes are the result of the quantization noise at the receiver that restrict the α from going to zero as $\rho=\infty$. For linear receivers ($b=\infty$), perfect channel estimation can be obtained as $\rho=\infty$, and therefore there are no asymptotes. At low SNR, the slopes of the curves are similar to each other because the additive (thermal) noise dominates the quantization noise, and the effect of quantization can be treated as degradation in SNR that depends on b.

IV. Discussion and Conclusion

We provided a method to compute the mutual information lower bound that can be used to analyze the effect of training in models with unknown parameters. For large-scale systems, we showed that this bound could be calculated using the difference between two derivatives of a conditional entropy function. This method was applied to a large-scale quantized MIMO system, where we derived the mutual information and studied how quantization and training influence one another. Several training results were derived that were unique to quantization effects at the transmitter and the receiver.

We believe that the analysis shown in Theorem 3 for the model (25) can be generalized beyond Gaussian channels and to other nonlinear functions $f(\cdot)$. In particular, since a quantizer with sufficiently high resolution and number of

levels can be used to approximate a well-behaved monotonic function, it is conceivable that the theorem can readily be adapted to any monotonic function. We view this as a possible avenue for future work.

In addition, we believe that Assumption A2 is likely to be superfluous for common system models, such as when the distribution on $\mathbf{x}_t$ is *iid* through the training and data phases, and the transition probabilities can be written as a product as in Assumption A1. However, we have not yet characterized for which models A2 is automatically satisfied without additional assumptions on $\mathcal{H}'$, and think that this would be an interesting research topic for further work.

Appendix A
Proof of Theorem 1

We first show the derivative relationship between $\mathcal{H}'(\mathcal{Y}_\varepsilon)$ and $\mathcal{H}(\mathcal{Y}_\varepsilon)$ defined below, and then generalize to the conditional entropies which directly lead to the conclusion (17). Define

$$\mathcal{H}'(\mathcal{Y}_\varepsilon) = \lim_{T\to\infty} \frac{1}{N} H(\mathbf{y}_{\lceil \varepsilon\tau T\rceil+1}|\mathbf{y}^{\lceil \varepsilon\tau T\rceil}), \tag{49}$$

$$\mathcal{H}(\mathcal{Y}_\varepsilon) = \lim_{T\to\infty} \frac{1}{N\tau T} H(\mathbf{y}^{\lceil \varepsilon\tau T\rceil}), \tag{50}$$

which can be considered as $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ and $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ with $\delta = 0$. For mathematical rigorousness, we keep the $\lceil\cdot\rceil$ notation here. We show that, under some conditions, $\mathcal{H}'(\mathcal{Y}_\varepsilon)$ is the derivative of $\mathcal{H}(\mathcal{Y}_\varepsilon)$.

Lemma 1: Suppose there exists a $\kappa > 0$ so that $H(\mathbf{y}_{t+1}|\mathbf{y}^t)$ is monotonic in t when $t \in [\lfloor(\varepsilon-\kappa)\tau T\rfloor, \lceil(\varepsilon+\kappa)\tau T\rceil]$ as $T \to \infty$.

If $\mathcal{H}(\mathcal{Y}_\varepsilon)$ and its derivative with respect to ε exist, we have

$$\mathcal{H}'(\mathcal{Y}_\varepsilon) = \frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon)}{\partial \varepsilon}. \tag{51}$$

If both $\mathcal{H}(\mathcal{Y}_\varepsilon)$ and $\mathcal{H}'(\mathcal{Y}_\varepsilon)$ exist, we have

$$\mathcal{H}(\mathcal{Y}_\varepsilon) = \int_0^\varepsilon \mathcal{H}'(\mathcal{Y}_u)du. \tag{52}$$

Proof: Equation (52) is an integral equivalent of (51) and we only prove (51) for simplicity. Without loss of generality, we assume that $H(\mathbf{y}_{t+1}|\mathbf{y}^t)$ is monotonically decreasing. Using the definition of $\mathcal{H}(\mathcal{Y}_\varepsilon)$ in (50), we have

$$\begin{aligned}
&\frac{1}{\kappa}(\mathcal{H}(\mathcal{Y}_{\varepsilon+\kappa}) - \mathcal{H}(\mathcal{Y}_\varepsilon)) \\
&= \lim_{T\to\infty} \frac{H(\mathbf{y}^{\lceil(\varepsilon+\kappa)\tau T\rceil}) - H(\mathbf{y}^{\lceil\varepsilon\tau T\rceil})}{\kappa N\tau T} \\
&= \lim_{T\to\infty} \frac{\sum_{t=\lceil\varepsilon\tau T\rceil+1}^{\lceil(\varepsilon+\kappa)\tau T\rceil} H(\mathbf{y}_t|\mathbf{y}^{t-1})}{\kappa N\tau T} \\
&\le \lim_{T\to\infty} \frac{(\lceil(\varepsilon+\kappa)\tau T\rceil - \lceil\varepsilon\tau T\rceil)}{\kappa N\tau T} \\
&\quad \times H(\mathbf{y}_{\lceil\varepsilon\tau T\rceil+1}|\mathbf{y}^{\lceil\varepsilon\tau T\rceil}) \\
&= \lim_{T\to\infty} \frac{1}{N} H(\mathbf{y}_{\lceil\varepsilon\tau T\rceil+1}|\mathbf{y}^{\lceil\varepsilon\tau T\rceil}).
\end{aligned} \tag{53}$$

Similarly to (53), we also have

$$\lim_{T\to\infty} \frac{1}{N} H(\mathbf{y}_{\lceil\varepsilon\tau T\rceil+1}|\mathbf{y}^{\lceil\varepsilon\tau T\rceil}) \le \frac{1}{\kappa}(\mathcal{H}(\mathcal{Y}_\varepsilon) - \mathcal{H}(\mathcal{Y}_{\varepsilon-\kappa})). \tag{54}$$

Let $\kappa \searrow 0$ in both (53) and (54); because we assume that the derivative of $\mathcal{H}(\mathcal{Y}_\varepsilon)$ exists, these limits both equal this derivative. Then, the definition of $\mathcal{H}'(\mathcal{Y}_\varepsilon)$ in (49) yields (51). □

Lemma 1 is a consequence of the entropy chain rule and letting an infinite sum converge to an integral (standard Riemann sum approximation). Such an analysis has also been used in the context of computing mutual information; for example [27]–[31]. Lemma 1 can be generalized to include conditioning on $\mathcal{X}$, thus leading to the following theorem, provided that $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$ and its derivative with respect to ε exist.

Theorem 4: Assume A1 holds. For $\varepsilon > 1$,

$$\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}) = \frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})}{\partial \varepsilon}, \tag{55}$$

$$\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+}) = \frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\varepsilon)}{\partial \varepsilon}. \tag{56}$$

Proof: Under A1, for all $\delta \ge 1$, we have

$$\begin{aligned}
H(\mathbf{y}_{t+1}|\mathbf{x}^{\lceil\delta\tau T\rceil}, \mathbf{y}^t) &\le H(\mathbf{y}_{t+1}|\mathbf{x}^{\lceil\delta\tau T\rceil}, \mathbf{y}^{t-1}) \\
&= H(\mathbf{y}_t|\mathbf{x}^{\lceil\delta\tau T\rceil}, \mathbf{y}^{t-1}),
\end{aligned} \tag{57}$$

when $\lceil\tau T\rceil + 1 \le t \le \lceil\delta\tau T\rceil - 1$ or $t \ge \lceil\delta\tau T\rceil + 1$. Here, we use that the input is *iid* and the system is memoryless and time invariant; the inequality follows from the fact that conditioning reduces entropy. Therefore, $\forall\kappa \in (0, \varepsilon-1)$, $H(\mathbf{y}_{t+1}|\mathbf{x}^{\lceil\tau T\rceil}, \mathbf{y}^t)$ is monotonically decreasing in t for $t \in [\lfloor(\varepsilon-\kappa)\tau T\rfloor, \lceil(\varepsilon+\kappa)\tau T\rceil]$ when $\tau T > \frac{2}{\varepsilon-1-\kappa}$. Then, Lemma 1 yields (55).

Also, for all $\delta > \varepsilon > 1$ there exists a $\kappa \in (0, \min(\varepsilon-1, \delta-\varepsilon))$, such that $H(\mathbf{y}_{t+1}|\mathbf{x}^{\lceil\delta\tau T\rceil}, \mathbf{y}^t)$ is monotonically decreasing in t for $t \in [\lfloor(\varepsilon-\kappa)\tau T\rfloor, \lceil(\varepsilon+\kappa)\tau T\rceil]$ when $\tau T > \max(\frac{3}{\varepsilon-1-\kappa}, \frac{2}{\delta-\varepsilon-\kappa})$. Then, Lemma 1 yields

$$\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)}{\partial \varepsilon}. \tag{58}$$

Assumption A1 yields

$$\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+}), \tag{59}$$

where $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+})$ is defined in (10), and

$$\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\varepsilon). \tag{60}$$

Therefore, (58) becomes (56). □

Theorem 4 can now be used to finish the proof of Theorem 1. By (14) and Assumptions A1–2,

$$\begin{aligned}
\mathcal{I}'(\mathcal{X};\mathcal{Y}) &= \mathcal{H}'(\mathcal{Y}|\mathcal{X}) - \mathcal{H}'(\mathcal{Y}|\mathcal{X}_+) \\
&= \lim_{\varepsilon\searrow 1} \mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}) - \lim_{\varepsilon\searrow 1} \mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+}),
\end{aligned}$$

and together with Theorem 4, we have (17).

Assumption A1 is often met in a wireless communication system, while whether Assumption A2 holds is not necessarily obvious. We provide the following corollary to help check Assumption A2.

Corollary 1: Assume A1 holds. (a) If $\mathbf{x}_t$ are iid for all t, then for all $\varepsilon, \delta > 0$ and $\varepsilon \ne \delta$,

$$\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta) = \frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)}{\partial \varepsilon}, \tag{61}$$

$$\mathcal{H}'(\mathcal{Y}|\mathcal{X}_+) = \left.\frac{\partial \mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=1}. \tag{62}$$

(b) If

$$\lim_{\varepsilon\searrow 1}\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X})=\lim_{\delta\nearrow 1}\mathcal{H}'(\mathcal{Y}|\mathcal{X}_\delta), \tag{63}$$

then Assumption A2 (12) *is met.*

Proof: (a) If $\mathbf{x}_t$ are *iid* for all t, then (57) is valid for all $t\le\lceil\delta\tau T\rceil-1$ or $t\ge\lceil\delta\tau T\rceil+1$. Therefore, Lemma 1 yields (61). By taking $\varepsilon=1$ and $\delta>1$, (61), (59), and (60) then yield (62).

(b) For $t\ge\lceil\tau T\rceil+1$, we have

$$H(\mathbf{y}_{t+1}|\mathbf{x}^{\lceil\tau T\rceil},\mathbf{y}^t)\le H(\mathbf{y}_t|\mathbf{x}^{\lceil\tau T\rceil},\mathbf{y}^{t-1}).$$

Therefore,

$$\lim_{\varepsilon\searrow 1}\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X})\le\mathcal{H}'(\mathcal{Y}|\mathcal{X}).$$

Conditioning to reduce entropy again yields

$$H(\mathbf{y}_{\lceil\tau T\rceil+1}|\mathbf{x}^{\lceil\tau T\rceil},\mathbf{y}^{\lceil\tau T\rceil})\le H(\mathbf{y}_{\lceil\tau T\rceil+1}|\mathbf{x}^{\lceil\delta\tau T\rceil},\mathbf{y}^{\lceil\tau T\rceil}),$$

for any $\delta<1$ and therefore,

$$\mathcal{H}'(\mathcal{Y}|\mathcal{X})\le\lim_{\delta\nearrow 1}\mathcal{H}'(\mathcal{Y}|\mathcal{X}_\delta).$$

Equation (63) then implies A2 (12).

□

APPENDIX B
PROOF OF THEOREM 3

A. Steps to Compute (37)

We now show the derivation of the expressions of $\mathcal{I}'(\mathcal{X};\mathcal{Y}),\mathcal{H}'(\mathcal{Y}|\mathcal{X}_+)$, and $\mathcal{H}'(\mathcal{Y}|\mathcal{X})$ by using theorems developed in Section II, which can be computed from a single entropy $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})=\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_{\delta=1})$, where $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)=\lim_{T\to\infty}\frac{1}{NTT}H(\mathbf{y}^{\lceil\varepsilon\tau T\rceil}|\mathbf{x}^{\lceil\delta\tau T\rceil})$. $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ can be obtained from a quantity called asymptotic free entropy $\mathcal{F}$ through

$$\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})=-\frac{\mathcal{F}(\tau\beta,(\varepsilon-1)\tau\beta)}{\alpha\tau\beta}, \tag{64}$$

where $\mathcal{F}(\beta_\tau,\beta_{\rm d})$ is defined as $\mathcal{F}(\beta_\tau,\beta_{\rm d})=\lim_{M\to\infty}\frac{1}{M^2}\mathbb{E}\log_2 p(\mathbf{y}^{\lceil(\beta_\tau+\beta_{\rm d})M\rceil}|\mathbf{x}^{\lceil\beta_\tau M\rceil})$, which has been computed as (44) in [12], and is continuous in β_τ and $\beta_{\rm d}$.

Following the steps in Section II-D:

1) From the model (25), it is clear that Assumption A1 is met, i.e.

$$p(\mathbf{y}^T|\mathbf{x}^T,G)=\prod_{t=1}^{T}p(\mathbf{y}_t|\mathbf{x}_t,G),$$

$$p(\mathbf{x}^T)=p(\mathbf{x}^{\tau T})\prod_{t=\tau T+1}^{T}p(\mathbf{x}_t),$$

where $p(\mathbf{y}_t|\mathbf{x}_t,G)$ is a time-invariant conditional distribution for all $t=1,2,\ldots,T$, and $p(\mathbf{x}_t)$ is a time-invariant distribution for all $t=\tau T+1,\tau T+2,\ldots,T$. The dimension of G depends on the blocklength T through (3). Furthermore, the input $\mathbf{x}_t$ are *iid* for all t. Then, we can express the system by a set of distributions defined as

$$\mathcal{P}(T,\tau)=\{p(\mathbf{y}|\mathbf{x},G),p(G),p(\mathbf{x})\}, \tag{65}$$

which are the conditional distribution of the system, the distribution of G, and the input distribution.

2) For given α and β, we define $F(\tau,\varepsilon)=\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$, where the entropy $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X})$ is computed through $\mathcal{P}(T,\tau)$ defined in (65). Then, (64) yields

$$F(\tau,\varepsilon)=-\frac{\mathcal{F}(\tau\beta,(\varepsilon-1)\tau\beta)}{\alpha\tau\beta}, \tag{66}$$

3) Theorem 2 then yields $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)=u\cdot F\left(u\tau,\frac{\varepsilon-u}{\delta}+1\right)$ for all $\varepsilon,\delta>0$, where $u=\min(\delta,\varepsilon)$. Then, (66) yields $\mathcal{H}(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)=-\frac{1}{\alpha\tau\beta}\mathcal{F}(u\tau\beta,(\varepsilon-u)\tau\beta)$.

4) Corollary 1(a) in Appendix A then yields

$$\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)=\begin{cases}\dfrac{-1}{\alpha\tau\beta}\dfrac{\partial}{\partial\varepsilon}\mathcal{F}(\varepsilon\tau\beta,0), & \varepsilon<\delta;\\ \dfrac{-1}{\alpha\tau\beta}\dfrac{\partial}{\partial\varepsilon}\mathcal{F}(\delta\tau\beta,(\varepsilon-\delta)\tau\beta), & \varepsilon>\delta.\end{cases}$$

for all $\varepsilon,\delta>0$. Since $\mathbf{x}_t$ is independent of $(\mathbf{x}_k,\mathbf{y}_k)$ when $t\ne k$, we have $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)=\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+})$, for all $\delta>\varepsilon>0$.

5) Using $\mathcal{F}(\cdot,\cdot)$ in [12] and Corollary 1(a), we can verify that

$$\mathcal{H}'(\mathcal{Y}|\mathcal{X}_+)=\lim_{\varepsilon\searrow 1}\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+}),$$

and

$$\mathcal{H}'(\mathcal{Y}|\mathcal{X}_+)=-\frac{1}{\alpha\tau\beta}\frac{\partial}{\partial\varepsilon}\mathcal{F}(\varepsilon\tau\beta,0)|_{\varepsilon=1}. \tag{67}$$

Moreover, using $\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_\delta)$, we have

$$\lim_{\varepsilon\searrow 1}\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X})=\lim_{\delta\nearrow 1}\mathcal{H}'(\mathcal{Y}|\mathcal{X}_\delta).$$

Thus, Assumption A2 is met via Corollary 1(b), i.e.

$$\text{A2:}\quad \mathcal{H}'(\mathcal{Y}|\mathcal{X}_+)=\lim_{\varepsilon\searrow 1}\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}_{\varepsilon^+}),$$
$$\mathcal{H}'(\mathcal{Y}|\mathcal{X})=\lim_{\varepsilon\searrow 1}\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X}),$$

and therefore,

$$\begin{aligned}\mathcal{H}'(\mathcal{Y}|\mathcal{X})&=\lim_{\varepsilon\searrow 1}\mathcal{H}'(\mathcal{Y}_\varepsilon|\mathcal{X})\\&=-\frac{1}{\alpha\tau\beta}\lim_{\varepsilon\searrow 1}\frac{\partial}{\partial\varepsilon}\mathcal{F}(\tau\beta,(\varepsilon-1)\tau\beta).\end{aligned}$$

6) Finally, Theorem 1 yields

$$\begin{aligned}\mathcal{I}'(\mathcal{X};\mathcal{Y})=&\frac{-1}{\alpha\tau\beta}\lim_{\varepsilon\searrow 1}\frac{\partial}{\partial\varepsilon}\mathcal{F}(\tau\beta,(\varepsilon-1)\tau\beta)\\&+\frac{1}{\alpha\tau\beta}\frac{\partial}{\partial\varepsilon}\mathcal{F}(\varepsilon\tau\beta,0)|_{\varepsilon=1}.\end{aligned} \tag{68}$$

By using the expression of $\mathcal{F}$ in [12], together with (68), we can derive (37). Specifically, converting the expression of $\mathcal{F}$ (44) in [12] to expressions with our notations, we get

$$\begin{aligned}&\mathcal{F}(\beta_\tau,\beta_{\rm d})\\&=\alpha\left[-\beta_\tau\Omega(\rho q_g,\sigma^2+\rho-\rho q_g)\right.\\&\quad\left.-\beta_{\rm d}\Omega(\rho q_g q_x,\sigma^2+\rho-\rho q_g q_x)\right]\\&\quad-\alpha I_{\rm AWGN}(\lambda_g,p_{\tilde g}(\cdot))-\beta_{\rm d}I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot))\\&\quad+\alpha\frac{(1-q_g)\lambda_g}{\ln 2}+\beta_{\rm d}\frac{(1-q_x)\lambda_x}{\ln 2},\end{aligned}$$

where $I_{\rm AWGN}(\cdot,\cdot)$, $\mathcal{E}(\cdot,\cdot)$, $\Omega(\cdot,\cdot)$, and $\chi(\cdot,\cdot)$ are defined in (28), (29), (32) and (33), and $(q_g,\lambda_g,q_x,\lambda_x)$ are the solutions of (38) and (39). Thus, the first term in (68) becomes

$$\begin{aligned}&-\frac{1}{\alpha\tau\beta}\lim_{\varepsilon\searrow 1}\frac{\partial}{\partial\varepsilon}\mathcal{F}(\tau\beta,(\varepsilon-1)\tau\beta)\\&=-\frac{1}{\alpha\tau\beta}\lim_{\varepsilon\searrow 1}\frac{\partial}{\partial\varepsilon}\Big\{\alpha\left[-\tau\beta\Omega(\rho q_g,\sigma^2+\rho-\rho q_g)\right.\\&\quad\left.-(\varepsilon-1)\tau\beta\Omega(\rho q_g q_x,\sigma^2+\rho-\rho q_g q_x)\right]\\&\quad-\alpha I_{\rm AWGN}(\lambda_g,p_{\tilde g}(\cdot))-(\varepsilon-1)\tau\beta I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot))\\&\quad+\frac{\alpha(1-q_g)\lambda_g}{\ln 2}+\frac{(\varepsilon-1)\tau\beta(1-q_x)\lambda_x}{\ln 2}\Big\},\end{aligned}$$

and the second term in (68) becomes

$$\begin{aligned}&\frac{1}{\alpha\tau\beta}\frac{\partial}{\partial\varepsilon}\mathcal{F}(\varepsilon\tau\beta,0)|_{\varepsilon=1}\\&=\frac{1}{\alpha\tau\beta}\frac{\partial}{\partial\varepsilon}\Big\{-\alpha\varepsilon\tau\beta\Omega(\rho q_g,\sigma^2+\rho-\rho q_g)\\&\quad-\alpha I_{\rm AWGN}(\lambda_g,p_{\tilde g}(\cdot))+\alpha\frac{(1-q_g)\lambda_g}{\ln 2}\Big\}|_{\varepsilon=1}.\end{aligned}$$

Then, by combining the two equations above, we can get (37).

APPENDIX C
$\mathcal{R}_{\rm opt}$ AND $\tau_{\rm opt}$ FOR LARGE α

To prove (40)–(42), we analyze $b=\infty$ and $b=1$ separately.

A. Large α With $b=\infty$

$\mathcal{I}'(\mathcal{X};\mathcal{Y})$ can be computed by following the steps in Section II-D. When $b=\infty$, (37) yields

$$\begin{aligned}\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})=\frac{1}{\ln 2}\Big(\alpha\ln\Big(1+\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x\Big)-\lambda_x\mathcal{E}_x\Big)\\+I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot)),\end{aligned}\tag{69}$$

where $p_{\tilde x}(x)$ is the distribution of real/imaginary part of elements of $\mathbf{x}_t$, $\bar\rho,\bar\sigma^2$ are computed from

$$\bar\rho=\rho q_g,\quad \bar\sigma^2=\sigma^2+\rho(1-q_g),\tag{70}$$

with q_g being the solution of (38), and $\mathcal{E}_x,\lambda_x$ are the solution of (39), which is

$$\lambda_x=\frac{\alpha\frac{\bar\rho}{\bar\sigma^2}}{1+\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x},\quad \mathcal{E}_x=\mathcal{E}(\lambda_x,p_{\tilde x}(\cdot)).\tag{71}$$

Then, (69) becomes

$$\begin{aligned}\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})=\frac{\alpha}{\ln 2}\ln\Big(1+\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x\Big)\\-\frac{\alpha}{\ln 2}\frac{\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x}{1+\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x}+I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot)).\end{aligned}\tag{72}$$

The mean-square error of the MMSE estimate is $\mathcal{E}(\lambda_x,p_{\tilde x}(\cdot))$ defined in (29), which is upper-bounded by the mean-square error of the LMMSE estimate, therefore we get

$$0\le\mathcal{E}(\lambda_x,p_{\tilde x}(\cdot))\le\frac{1}{1+\lambda_x},\tag{73}$$

$$\mathcal{E}_x<\frac{1}{\lambda_x},\quad 0\le\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x\le\frac{1}{\alpha-1}.\tag{74}$$

Because $\ln(1+w)-\frac{w}{1+w}$ is monotonically increasing in w for $w\ge 0$, (74) yields $0\le\alpha\left(\ln\left(1+\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x\right)-\frac{\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x}{1+\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x}\right)\le\alpha(\frac{1}{\alpha-1}-\frac{1}{\alpha})=\frac{1}{\alpha-1}$. Therefore, (72) yields $I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot))\le\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})\le\frac{1}{\ln 2(\alpha-1)}+I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot))$, thus

$$\lim_{\alpha\to\infty}\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})=\lim_{\alpha\to\infty}I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot)).\tag{75}$$

As $\alpha\to\infty$, (71) and (74) imply that $\lambda_x\to\infty$, and therefore $I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot))\to H(x)$ which is the entropy of x. For $2^{2a}-$QAM moduation at the transmitter generated by a-bit DAC's, we have $\lim_{\alpha\to\infty}\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})=H(x)=2a$ for any finite a, and (26), (5) then yield

$$\lim_{\alpha\to\infty}\mathcal{R}_{\rm opt}=2a,\quad\lim_{\alpha\to\infty}\tau_{\rm opt}=0.\tag{76}$$

This shows (40).

When $a=1$, $\mathcal{E}(\lambda_x,p_{\tilde x}(\cdot))$ is upper-bounded by the MSE obtained through a hard decision, or

$$\mathcal{E}_x=\mathcal{E}(\lambda_x,p_{\tilde x}(\cdot))\le 4Q(\sqrt{\lambda_x}).\tag{77}$$

For large α, (71) and (77) imply that $\mathcal{E}_x$ decays exponentially to zero, and therefore

$$\lambda_x\approx\alpha\bar\rho/\bar\sigma^2,\tag{78}$$

$$\begin{aligned}\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})=\frac{\alpha}{2\ln 2}\Big(\frac{\bar\rho}{\bar\sigma^2}\mathcal{E}_x\Big)^2\\+I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot))+o(\mathcal{E}_x^2).\end{aligned}\tag{79}$$

Equation (76) implies that $\tau_{\rm opt}$ is small when α is large, and therefore, in the following approximations, we only keep the dominant terms in τ. Equations (38) and (70) yield $q_g\approx\frac{\rho}{1+\rho}\tau\beta$ and $\frac{\bar\rho}{\bar\sigma^2}\approx\left(\frac{\rho}{1+\rho}\right)^2\tau\beta$, and (78) then yields $\lambda_x\approx\rho^2\tau\beta\alpha/(1+\rho)^2$. It can be shown when $a=1$ that for some $\nu>0$, $\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})\approx 2-\nu e^{-\frac{\lambda_x}{2}}\sqrt{\lambda_x}$, and that therefore $\tau_{\rm opt}\approx\operatorname{argmax}_\tau(1-\tau)(2-\nu e^{-\frac{\nu_1\alpha\tau}{2}}\sqrt{\nu_1\alpha\tau})$ where $\nu_1=(\frac{\rho}{1+\rho})^2\beta$. Taking the derivative with respect to τ and setting it equal to zero produces (41).

B. Large α With $b=1$

We again compute $\mathcal{I}'(\mathcal{X};\mathcal{Y})$ by following the steps in Section II-D. When $b=1$, (37) yields

$$\begin{aligned}\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y})=4\alpha\int_{\mathcal{R}}[Q\left(\sqrt{\bar c}z\right)\log_2 Q\left(\sqrt{\bar c}z\right)\\-Q\left(Az\right)\log_2 Q\left(Az\right)]\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}dz\\-\frac{\lambda_x\mathcal{E}_x}{\ln 2}+I_{\rm AWGN}(\lambda_x,p_{\tilde x}(\cdot)),\end{aligned}\tag{80}$$

where

$$\bar c=\frac{\bar\rho}{\bar\sigma^2},\quad A=\sqrt{\frac{\bar c(1-\mathcal{E}_x)}{1+\bar c\mathcal{E}_x}},$$

$\bar\rho,\bar\sigma^2$ are computed from (70) which are not functions of α, and $(\mathcal{E}_x,\lambda_x)$ are the solution of (39), which can be expressed

as

$$\lambda_x = \int_{\mathcal{R}} dz \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \frac{\alpha\bar{c}}{\pi(1+\bar{c}\mathcal{E}_x)} \frac{e^{-A^2z^2}}{Q(Az)},$$
$$\mathcal{E}_x = \mathcal{E}(\lambda_x, p_{\tilde{x}}(\cdot)). \tag{81}$$

Since $0 < Q(Az) < 1$, (81) yields

$$\lambda_x \geq \frac{\alpha\bar{c}}{\pi(1+\bar{c}\mathcal{E}_x)} \int_{\mathcal{R}} dz \frac{e^{-\frac{1+2A^2}{2}z^2}}{\sqrt{2\pi}} = \frac{\alpha\bar{c}\sqrt{1+2A^2}}{\pi(1+\bar{c}\mathcal{E}_x)} \geq \frac{\bar{c}\alpha}{\pi(1+\bar{c})}. \tag{82}$$

Similar to (73), we have $0 \leq \mathcal{E}(\lambda_x, p_{\tilde{x}}(\cdot)) = \mathcal{E}_x \leq \frac{1}{1+\lambda_x}$. Then, (81) and (82) yield

$$0 \leq \mathcal{E}_x < \frac{1}{\lambda_x} \leq \frac{\pi(1+\bar{c})}{\bar{c}} \cdot \frac{1}{\alpha}.$$

Therefore, $\mathcal{E}_x$ becomes small for large α. A Taylor expansion of (80) obtains

$$\alpha\mathcal{I}'(\mathcal{X};\mathcal{Y}) = \frac{\alpha}{\pi \ln 2} \int_{\mathcal{R}} \frac{3A^2+2A-1}{(1+A^2)(1+(A+\epsilon)^2)} \times \frac{\epsilon^2 e^{-\frac{1+2A^2}{2}z^2}}{\sqrt{2\pi}Q(Az)} dz + I_{\mathrm{AWGN}}(\lambda_x, p_{\tilde{x}}(\cdot)) + O(\alpha\mathcal{E}_x^2),$$

where $\epsilon = \frac{(1+\bar{c})\bar{c}\mathcal{E}_x}{(\sqrt{\bar{c}}+A)(1+\bar{c}\mathcal{E}_x)}$. Thus,

$$\lim_{\alpha\to\infty} \alpha\mathcal{I}'(\mathcal{X};\mathcal{Y}) = \lim_{\alpha\to\infty} I_{\mathrm{AWGN}}(\lambda_x, p_{\tilde{x}}(\cdot)).$$

The remaining steps are similar to $b = \infty$ in Appendix C-A and are omitted.

Appendix D
Calculation of (46)

Since

$$\lim_{M\to\infty} \frac{1}{M} I(\mathbf{x};\bar{\mathbf{y}}|\bar{G}) = \lim_{M\to\infty} \frac{1}{M} H(\bar{\mathbf{y}}|\bar{G}) - \lim_{M\to\infty} \frac{1}{M} H(\bar{\mathbf{y}}|\bar{G},\mathbf{x}),$$

we compute the two terms separately as follows. When $\bar{G}$ is known, the entropy $\lim_{M\to\infty} \frac{1}{M} H(\bar{\mathbf{y}}|\bar{G})$ can be computed through

$$\lim_{M\to\infty} \frac{1}{M} H(\bar{\mathbf{y}}|\bar{G}) = \lim_{M\to\infty} \frac{-1}{\beta M^2} \mathbb{E} \log_2 p(\bar{\mathbf{y}}^{\lceil \beta M \rceil}|\bar{G}) = \alpha\Omega(\rho_{\mathrm{eff}}\bar{q}_x, 1+\rho_{\mathrm{eff}}-\rho_{\mathrm{eff}}\bar{q}_x) + I_{\mathrm{AWGN}}(\bar{\lambda}_x, p_{\tilde{x}}(\cdot)) - \frac{(1-\bar{q}_x)\bar{\lambda}_x}{\ln 2}, \tag{83}$$

where $\lim_{M\to\infty} \frac{1}{M^2} \mathbb{E} \log_2 p(\bar{\mathbf{y}}^{\lceil \beta M \rceil}|G)$ is available in [12] and $(\bar{q}_x, \bar{\lambda}_x)$ are the solutions of (47), and $\Omega(\cdot,\cdot)$ and $\chi(\cdot,\cdot)$ are defined in (36).

Therefore, (83) yields

$$\lim_{M\to\infty} \frac{1}{M} H(\bar{\mathbf{y}}|\bar{G}) = \alpha\Omega(\rho_{\mathrm{eff}}\bar{q}_x, 1+\rho_{\mathrm{eff}}-\rho_{\mathrm{eff}}\bar{q}_x) + I_{\mathrm{AWGN}}(\bar{\lambda}_x, p_{\tilde{x}}(\cdot)) + \frac{\bar{q}_x\bar{\lambda}_x - \bar{\lambda}_x}{\ln 2}. \tag{84}$$

Since the elements of $\bar{G}$ are *iid* $\mathcal{CN}(0,1)$, for any given $\mathbf{x}$, the elements of $\sqrt{\frac{\rho_{\mathrm{eff}}}{M}}\bar{G}\mathbf{x}$ are *iid* $\mathcal{CN}(0, \frac{\rho_{\mathrm{eff}}\mathbf{x}^{\mathsf{H}}\mathbf{x}}{M})$. Also, since the elements of $\mathbf{x}$ are *iid* with zero mean and unit variance, $\frac{\rho_{\mathrm{eff}}\mathbf{x}^{\mathsf{H}}\mathbf{x}}{M}$ converges to ρ_{eff} and the elements of $\sqrt{\frac{\rho_{\mathrm{eff}}}{M}}\bar{G}\mathbf{x}$ converge to *iid* $\mathcal{CN}(0,\rho_{\mathrm{eff}})$ as $M\to\infty$. Therefore,

$$\lim_{M\to\infty} \frac{1}{M} H(\bar{\mathbf{y}}|\bar{G},\mathbf{x}) = \alpha\Omega(\rho_{\mathrm{eff}},1). \tag{85}$$

Then, (84) and (85) yield

$$\begin{aligned}
&\lim_{M\to\infty} \frac{1}{M} I(\mathbf{x};\bar{\mathbf{y}}|\bar{G}) \\
&= \alpha\Omega(\rho_{\mathrm{eff}}\bar{q}_x, 1+\rho_{\mathrm{eff}}-\rho_{\mathrm{eff}}\bar{q}_x) - \alpha\Omega(\rho_{\mathrm{eff}},1) \\
&\quad + I_{\mathrm{AWGN}}(\bar{\lambda}_x, p_{\tilde{x}}(\cdot)) + \frac{\bar{q}_x\bar{\lambda}_x - \bar{\lambda}_x}{\ln 2} \\
&= \alpha \log_2(1+\rho_{\mathrm{eff}}-\rho_{\mathrm{eff}}\bar{q}_x) + I_{\mathrm{AWGN}}(\bar{\lambda}_x, p_{\tilde{x}}(\cdot)) \\
&\quad + \frac{\bar{q}_x\bar{\lambda}_x - \bar{\lambda}_x}{\ln 2},
\end{aligned}$$

where $(\bar{q}_x, \bar{\lambda}_x)$ are the solutions of (47), which finishes the computation of (46).

References

[1] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[2] R. Muharar, "Optimal power allocation and training duration for uplink multiuser massive MIMO systems with MMSE receivers," *IEEE Access*, vol. 8, pp. 23378–23390, 2020.

[3] K. Takeuchi, M. Vehkapera, T. Tanaka, and R. R. Müller, "Large-system analysis of joint channel and data estimation for MIMO DS-CDMA systems," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1385–1412, Mar. 2012.

[4] K. Takeuchi, R. R. Müller, M. Vehkaperä, and T. Tanaka, "On an achievable rate of large Rayleigh block-fading MIMO channels with, no CSI," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6517–6541, Oct. 2013.

[5] N. I. Miridakis and T. A. Tsiftsis, "On the joint impact of hardware impairments and imperfect CSI on successive decoding," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 4810–4822, Jun. 2016.

[6] Z. Sheng, H. D. Tuan, H. H. Nguyen, and M. Debbah, "Optimal training sequences for large-scale MIMO-OFDM systems," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3329–3343, Jul. 2017.

[7] Y. Li, C. Tao, L. Liu, A. Mezghani, and A. L. Swindlehurst, "How much training is needed in one-bit massive MIMO systems at low SNR?" in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[8] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Apr. 2017.

[9] J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. 216, 1952.

[10] C.-K. Wen, Y. Wu, K.-K. Wong, R. Schober, and P. Ting, "Performance limits of massive MIMO systems based on Bayes-optimal inference," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., 2015, pp. 1783–1788.

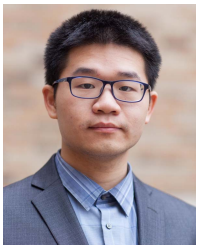
[11] C.-K. Wen, S. Jin, K.-K. Wong, C.-J. Wang, and G. Wu, "Joint cHANNEL-and-dATA estimation for large-MIMO systems with low-precision ADCs," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1237–1241.

[12] C. K. Wen, C. J. Wang, S. Jin, K. K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

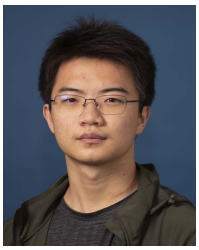
[13] T. Castellani and A. Cavagna, "Spin-glass theory for pedestrians," *J. Stat. Mech. Theory Exp.*, vol. 2005, no. 5, May 2005, Art. no. P05012.

[14] M. Mezard and A. Montanari, *Information, Physics, and Computation*. New York, NY, USA: Oxford Univ. Press, 2009.

[15] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning. Cambridge*. Cambridge, U.K.: Cambridge Univ. Press, 2001.
[16] M. Opper and W. Kinzel, "Statistical mechanics of generalization," in *Models of Neural Networks III* (Physics of Neural Networks), K. Schulten, E. Domany, and J. Leo van Hemmen, Eds. New York, NY, USA: Springer, 1996, ch. 5, pp. 151–209.
[17] T. Shinzato and Y. Kabashima, "Learning from correlated patterns by simple perceptrons," *J. Phys. A, Math. Theor.*, vol. 42, no. 1, pp. 015005-1–015005-12, 2008.
[18] S. Ha, K. Kang, J.-H. Oh, C. Kwon, and Y. Park, "Generalization in a perceptron with a sigmoid transfer function," in *Proc. Int. Conf. Neural Netw. (IJCNN)*, Nagoya, Japan, vol. 2, Oct. 1993, pp. 1723–1726.
[19] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
[20] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
[21] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. New York, NY, USA: Cambridge Univ. Press, 2016.
[22] E. Bjornson, L. Van der Perre, S. Buzzi, and E. G. Larsson, "Massive MIMO in sub-6 GHz and mmWave: Physical, practical, and use-case differences," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 100–108, Apr. 2019.
[23] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.
[24] K. Gao, J. N. Laneman, N. Estes, J. Chisum, and B. Hochwald, "Training for channel estimation in nonlinear multi-antenna transceivers," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, 2019, pp. 1–11.
[25] K. Gao, J. N. Laneman, N. J. Estes, J. Chisum, and B. Hochwald, "Channel estimation with one-bit transceivers in a Rayleigh environment," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.
[26] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, Jun. 2017.
[27] S. Shamai and S. Verdú, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1302–1327, May 2001.
[28] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1983–2010, May 2005.
[29] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.
[30] D. Guo and C.-C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 421–431, Apr. 2008.
[31] M. L. Honig, *Advances in Multiuser Detection*. Hoboken, NJ, USA: Wiley, 2009.



Kang Gao was born in Hubei, China. He received the B.S. degree in electrical engineering from the Huazhong University of Science and Technology, Hubei, in 2014, and the M.S. and Ph.D. degrees in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA. He joined Qualcomm Inc., San Diego, CA, USA, as a Senior Modem Systems Engineer in 2020. His research interests include millimeter-wave communication, nonlinear large-scale systems, and information theory.



Xiangbo Meng (Graduate Student Member, IEEE) received the B.S. degree in telecommunications from Northeastern University, Shenyang, China, in 2018, and the M.S. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 2020. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Notre Dame. His research interests include low-resolution wireless communication systems, spectrum confinement coding schemes, and information theory.

J. Nicholas Laneman (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT). He is the Director of SpectrumX—An NSF Spectrum Innovation Center, the Founding Director and currently the Co-Director of the Wireless Institute at the College of Engineering, and a Professor with the Department of Electrical Engineering, University of Notre Dame. He joined the Faculty in August 2002 after his Ph.D. degree. He is an author or coauthor of over 145 publications and is a co-inventor of eight U.S. patents. His research and teaching interests are in wireless systems design, radio spectrum access, technology standards and intellectual property, and regulatory policy. He has received the IEEE Kiyo Tomiyasu Award, the Presidential Early-Career Award for Scientists and Engineers (PECASE), and the NSF CAREER Award; and has been recognized twice by Thomson Reuters as an ISI Highly Cited Researcher.

Jonathan D. Chisum (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Colorado Boulder, Boulder, CO, USA, in 2011.

From 2012 to 2015, he was a Member of Technical Staff with the Lincoln Laboratory, Massachusetts Institute of Technology, in the Wideband Communications and Spectrum Operations groups. His work at the Lincoln Laboratory focused on millimeter-wave phased arrays, antennas, and transceiver design for electronic warfare applications. In 2015, he joined the Faculty of the University of Notre Dame, where he is currently an Associate Professor of electrical engineering. His research interests include millimeter-wave communications and spectrum sensing with an emphasis on low-power and low-cost technologies. His group focuses on gradient index (GRIN) lenses for low-power millimeter-wave beam-steering antennas, nonlinear (1-bit) radio architectures for highly efficient millimeter-wave communications and sensing, and the use of novel materials and micromagnetic effects for low-power analog signal processing and tunable antennas.

Dr. Chisum is a member of the American Physical Society and an elected Member of the U.S. National Committee (USNC) of the International Union of Radio Science's (URSI) Commission D (electronics and photonics). He is the Chair of USNC URSI Commission D and an Associate Editor of *IET Electronics Letters*.

Ralf Bendlin (Senior Member, IEEE) received the bachelor's and master's degrees from the Technical University of Munich, Munich, Germany, and the master's and Ph.D. degrees from the University of Notre Dame, South Bend, IN, USA, in electrical engineering and information technology. He is a Lead Member of Technical Staff at AT&T Labs, Austin, TX, USA. Before joining AT&T, he was a Senior Wireless Systems Architect with Intel Corporation and a Systems Engineer with Texas Instruments Inc. He has worked on algorithm development, performance prediction, optimization, systems architecture, spectrum research, and technology strategy for current and next-generation wireless networks; and has actively participated in the definition of several global communications standards. He was a Guest Editor of the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING (TCCN) and the Technical Program Chair of both the IEEE GLOBECOM 2019 Workshop on High-Dimensional, Low-Resolution Architectures for Power-Efficient Wireless Communications and the 2017 IEEE International Symposium on Dynamic Spectrum Access Networks.

Aditya Chopra (Senior Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Delhi, India, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from The University of Texas at Austin, Texas, USA, in 2008 and 2011, respectively. He has worked at Fastback Networks as a Communication Systems Engineer, National Instruments as a Senior Hardware Engineer, and AT&T Labs as a Principal Member of Technical Staff. In 2022, he joined Project Kuiper at Amazon as a Senior Communication Systems Engineer. His research interests include wireless physical layer optimization and prototyping of advanced wireless communication systems.

Bertrand M. Hochwald (Fellow, IEEE) was born in New York, NY, USA. He received the bachelor's degree from the Swarthmore College, Swarthmore, PA, USA, the M.S. degree in electrical engineering from Duke University, Durham, NC, USA, and the M.A. degree in statistics and the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA.

From 1986 to 1989, he was with the Department of Defense, Fort Meade, MD, USA. He was a Research Associate and a Visiting Assistant Professor with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL, USA. In 1996, he joined the Mathematics of Communications Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, USA, where he was a Distinguished Member of the Technical Staff. In 2005, he joined Beceem Communications, Santa Clara, CA, USA, as the Chief Scientist and the Vice-President of systems engineering. He served as a Consulting Professor of electrical engineering with Stanford University, Palo Alto, CA, USA. In 2011, he joined the University of Notre Dame, Notre Dame, IN, USA, as a Freimann Professor of electrical engineering. He has 46 patents and has co-invented several well-known multiple-antenna techniques, including a differential method, linear dispersion codes, and multi-user vector perturbation methods.

Dr. Hochwald currently serves on the technical advisory boards of wireless companies and is a fellow of the National Academy of Inventors. He received several achievement awards while employed at the Department of Defense and the Prize Teaching Fellowship at Yale University. He received the 2006 Stephen O. Rice Prize for the Best Paper published in the IEEE TRANSACTIONS ON COMMUNICATIONS. He has coauthored a paper that won the 2016 Best Paper Award by a young author in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS. He also won the 2018 H. A. Wheeler Prize Paper Award from the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION. His Ph.D. students have won various honors for their Ph.D. research, including the 2018 Paul Baran Young Scholar Award from the Marconi Society. He is listed as a Thomson Reuters Most Influential Scientific Mind in multiple years. He has served as an editor for several IEEE journals and has given plenary and invited talks on various aspects of signal processing and communications.